Matching of Matching-Graphs - A Novel Approach for Graph Classification

M. Fuchs, Institute of Computer Science, University of Bern, 3012 Bern, Switzerland
mathias.fuchs@inf.unibe.ch
K. Riesen, Institute of Informations Systems, University of Appl. Sci. Northwestern Switzerland, 4600 Olten, Switzerland
kaspar.riesen@fhnw.ch

UNIVERSITÄT BERN

Introduction

Graphs are recognized as versatile alternative to feature vectors. That is, graphs are used in diverse applications (protein function/structure prediction, signature verification or detection of Alzheimer's Disease). A large amount of graph based methods for pattern recognition have been proposed. *Graph edit distance* (GED) is one of the most flexible distance models available. GED generally offers more information than merely a dissimilarity score, namely the information of the objects actually match with each other (known as edit path).

Matching-Graphs

Given two graphs g_1 and g_2, the basic idea of GED is to transform g_1 into g_2 using some *edit operations* (e.g. *insertions*, *deletions*, and *substitutions* of both nodes and edges). The set of operations used to transform g_1 into g_2 is called *edit path*. For each edit path, two matching-graphs $m_{g_i \times g_j}$ and $m_{g_j \times g_i}$ can eventually be built (for the source and the target graph g_i and g_j, respectively). To this end, all nodes that are substituted are added to the matching-graphs. All nodes that are deleted or inserted are neither considered in the two matching-graphs. Nodes without adjacent nodes are removed from the resulting matching-graphs.
In Fig. 1 we can see an example of the source matching-graph created from the example edit path $\lambda = \{0 \to 0, 1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4, 5 \to \varepsilon, \varepsilon \to 5, \varepsilon \to 6, \varepsilon \to 7\}$.

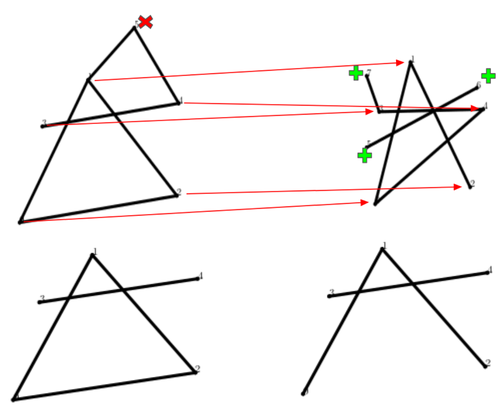

Figure 1: Visualization of an edit path and the resulting matching-graphs.

To handle the edges, we propose two different approaches: *pruned* and *unpruned* matching-graphs. For a pruned matching-graph if two nodes $u_1, u_2 \in V_i$ of a source graph g_i are substituted with nodes $v_1, v_2 \in V_j$ in a target graph g_j and there is an edge $(u_1, u_2) \in E_i$ available, (u_1, u_2) is actually included in the matching-graph $m_{g_i \times g_j}$ if, and only if, there is an edge (v_1, v_2) available in E_j. In an unpruned matching-graph it is included regardless whether or not edge (v_1, v_2) is available in E_j.

Classification using matching-graphs

The novel matching-graphs are employed in a distance based classification scenario. We define a distance measure that combines two distances with each other. A Visual illustration of this process can be seen in Fig. 2. The orange part $d_{BP}(g, g_i)$ is the distance information between a given test graph g and a training graph g_i, calculated using the BP approximation algorithm for GED. The green part $\min_{m \in \mathcal{M}_{\omega_l}} d_{BP}(g, m)$ is the minimum of all distances from g to all matching graphs m that are of the same class as g_i. We use $\alpha \in [0, 1]$ as a weighting parameter.

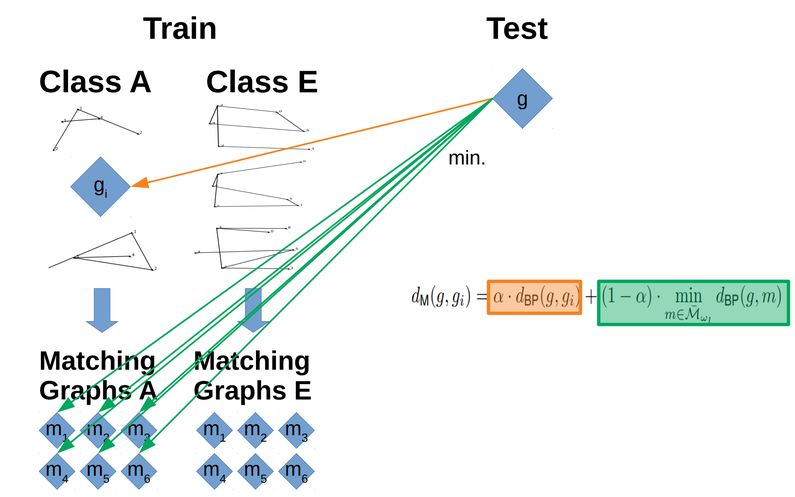

Figure 2: Example of the distance calculation using matching-graphs

Results

The reference system, denoted as d_{BP}, is based on the GED approximation algorithm BP in conjunction with the training graphs only. Our system, denoted as d_M, uses the same GED approximation yet makes use of the novel matching-graphs.

	k-NN(d_{BP})	k-NN(d_M)	
Data Set		Unpruned	Pruned
Letter	90.5	91.3	93.1 ∘
AIDS	99.0	99.7 ∘	99.7 ∘
Mutagenicity	70.6	70.0	70.5

Table 1: Results Table. On the left the baseline and on the right our approach. The ∘ indicates a statistically significant improvement over the reference system using a Z-Test at significance level $\alpha = 0.05$

We observe that the distance-based classification achieves better results with d_M rather than d_{BP} on two data sets (Letter and AIDS) and with both strategies (pruned and unpruned edges). On the Mutagenicity data set a slight - non statistically significant - deterioration is observed with our novel approach.
Hence, we conclude that the integration of matching-graphs can lead to a more accurate determination of a dissimilarity score using GED.